September 12, 2000

VETERINARY SERVICES MEMORANDUM NO. 800.96

Subject: Guidelines for Submitting Electronic Data Files for Statistical Analysis

To: Veterinary Biologics Licensees, Permittees, and Applicants

I. PURPOSE

This memorandum provides guidelines for the submission of supporting data in the form of electronic files to the Center for Veterinary Biologics-Licensing and Policy Development and supplements the Guidelines for Submission of Materials in Support of Licensure specified in Veterinary Biologics Memorandum 800.84.

II. CANCELLATION

This Memorandum cancels Veterinary Services Memorandum No. 800.96 dated November 22, 1999.

III. BACKGROUND

Product license applications must be supported by data according to 9 CFR 102.3(b) (2) (ii). Increasingly, license applicants use electronic files to submit data supporting applications for United States Veterinary Biological Product Licenses. Typically, applicants present these data for visual impact within the text of the report. Such arrangement, however, usually results in data that are unsuitable for analysis by computer. To facilitate the timely and efficient review of data, APHIS recommends that data submitted in electronic form adhere to the guidelines set forth in Veterinary Services Memorandum No. 800.84, section IV, Presentation of Data, and to the guidelines described below.

IV. DATA GUIDELINES

Applicants should submit data to CVB - LPD without alteration in a format readily amenable to analysis. Applicants should include all recorded data for each reported outcome, and indicate if the analysis in the report is based on a subset of the recorded data. If a subset of the data is used, applicants should justify this use in the accompanying text.

A. Recording the Raw Data

While analysis may, at times, be conducted on simplified forms of the data, applicants should submit the complete data in its raw unaltered form.

1. *Form Observed* - Record the data in the form observed before transformation, rounding, or other manipulation.

2. *Precision Observed* - Record the data with the precision observed. For example, record the number of animals as 7, not 7.0. In general, the number of digits should be neither too many (false precision) nor too few (rounding error.)

3. *Without Reduction* - Record the data without reduction, such as by summing or averaging.

4. *Categorical Data* - submit categorical data before collapsing categories.

a. Continuous outcomes which have been categorized should be submitted in raw form. For example, if white blood cell (WBC) counts have been recorded and then dichotomized into low (leukopenia) and normal, submit the raw WBC counts as the outcome in the data table.

b. Multiple categories which have been collapsed into fewer categories should be submitted in the original form.

B. Organization of Raw Data

1. *Arrayal* - Array the data as a table.

2. *Content* - The table should contain only raw data. Do not enter derived information such as averages or totals. Do not include graphic enhancements such as lines or shading.

3. *Headings* - Assign a single heading to each column and, if necessary, to each row.

4. *Entries* - Include a separate entry in each cell, even when entries for adjacent cells are identical.

5. *Unavailable Data* - Designate unavailable data by entering a single non-numeric code (such as "NA" or ".") in the appropriate cell. Do not leave the cell blank.

V. DATA FORMAT

A. Tabular Formats.

Arrange the data tables according to one of the following two formats.

1. *Univariate Table* - In a univariate table, each observed response, such as a serum titer, occupies its own row along with the values of explanatory variables, such as body weight, and factors, such as vaccination status, for that observation. The univariate table is the most commonly used format for data analysis and applicants should use it whenever possible. The construction of the table should completely partition the data so that:

a. Each row of the table contains a separate observation (the entries are mutually exclusive.)

b. Each observation is present in one row of the table (the entries are collectively exhaustive.)

2. *Multivariate Table* - A response may take the form of a cluster of related observations rather than a single measurement. Examples of observation clusters include a series of body temperatures measured daily on one animal over a week, a series of optical density measurements made over a sequence of dilutions, or a group of potency values concurrently measured for several fractions of a multivalent vaccine. Under certain conditions, clustered observations may be arrayed in a multivariate table which, compared to a univariate table, saves space as well as visually highlights the clustered nature of the observations. A multivariate table should have the following features:

a. Each row contains a cluster of related response observations.

b. Values of the primary explanatory variable are given by the headings of the columns containing the response cluster. Thus, the headings for a set of body temperatures recorded on a single subject may indicate that the measurements were taken on days 1 through 7. This set of values is common to all the clusters of observations.

c. Additional columns in the table may contain other explanatory variables, such as the subjects' age. In each row, the value of such a variable is common to the entire cluster of observations on that row.

d. If the values of an explanatory variable differ both within and between clusters, the multivariate format is not suitable; and the univariate format should be used.

B. Format for Categorical Data

If there are only two factors including the outcome, the data may be submitted in the form of a contingency table or as a frequency table. If there are more than two factors, a frequency table should be used.

1. *Contingency Table* - Contingency tables which cross-classify response counts are a useful visual display for categorical frequency data. In a contingency table, each cell shows the frequency of a unique combination of the categories which are identified on the margins of the table.

2. *Frequency Table* - A frequency table is a univariate table which has been collapsed over identical category combinations. The explanatory variables in each row consist of a particular combination of the categories of the factors. The outcome in each row is the number of obervations at that particular combination. To completely partition the data, the table should include a row for each possible unique combination of the categories of every factor. If any such combinations are omitted from the table, CVB will assume their frequencies are zero. Denominator data, such as the number of doses used, may be included where appropriate.

VI. SUBMITTING DATA AND ANALYSES

A. Data File

Submit data tables in electronic files. Electronic text files are appropriate for most data sets of the type normally submitted to the Center for Veterinary Biologics. They have the advantages of being both portable and readily perused visually. Data sets that are too large to be easily handled in a tab delimited text file may be submitted in a suitable data file format such as those generated by major statistical or database software. Such files should follow the guidelines in this document. It is advisable to accompany each file with a description of its contents, including the size and nature of the data set, missing data code, and other practical information.

B. Data Description

Include a separate file indexing the data files and describing their contents. The description should state each variable's name (eg. Wbc), Nature (eg. White blood cell count) and measurement unit (eg. 106 per ml). It should define all codes (eg. A = vaccine, B = placebo, including the missing data code (eg. Missing data indicated by NA) if applicable. It should state the number of observations and variables in the data set.

CC. Data Analysis

Identify software used in the analysis. Review and processing of the submission will be expedited by including the software output and the programming code or command sequence used to generate the output. For software that is entirely menu driven, include a session log. For spreadsheets, include cell formulae.

VII. EXAMPLES

The enclosed examples illustrate an appropriate format for several common data types.

/s/

Alfonso Torres
Deputy Administrator
Veterinary Services

Enclosure

Example 1 – ELISA

Incorrect

Test 1 of 3

New Vaccine								
	With Adjuvant				Without Adjuvant			
Dilution	1	2	3	Average	4	5	6	Average
undiluted	0.8347	0.7937	0.8337	0.8207	0.8097	0.6797	0.8077	0.7657
1:2	0.8427	0.7967	0.8527	0.8307	0.6917	0.6697	0.7047	0.6887
1:4	0.7187	0.6957	0.6827	0.699	0.4747	0.6067	0.5577	0.5464
1:8	0.5907	0.6787	0.5977	0.6224	0.5117	0.7747	0.6857	0.6574
1:16	0.2517	0.2297	0.2497	0.2437	0.1677	0.1717	0.1627	0.1674
1:32	0.08267	0.09367	0.09767	0.0913	0.005667	0.003667	0.03867	0.016
1:64	-0.00233	-0.0253	-0.0243	-0.0173	-0.0583	-0.0303	-0.0593	-0.0493
Old Vaccine								
	With Adjuvant				Without Adjuvant			
Dilution	7	8	9	Average	10	11	12	Average
undiluted	1.2567	1.2257	1.1457	1.2094	1.2757	1.4687	1.3437	1.3627
1:2	0.6867	0.7477	0.7257	0.72	0.9317	1.0127	1.0617	1.002
1:4	0.3457	0.3587	0.3487	0.351	0.4187	0.6307	0.6867	0.5787
1:8	0.1687	0.1427	0.1157	0.1424	0.1917	0.3777	0.4167	0.3287
1:16	0.02567	0.002667	0.01267	0.0137	0.04767	0.1387	0.1377	0.108
1:32	-0.0833	-0.0743	-0.0763	-0.078	-0.0523	-0.0273	-0.0173	-0.0323
1:64	-0.117	-0.102	-0.0993	-0.1061	0.03067	-0.0613	-0.0353	-0.022

This display cannot serve as a raw data table because there are more than one set of observations per row (with and without adjuvant), items that are not raw data are included (averages), explanatory variables appear neither in the table nor as column headings (test appears in the title; vaccine and adjuvant are in subtable headings), and there are unnecessary column headings (the numbers 1 through 12).

Correct – Univariate Format

To array these data in univariate format, stack each column under the preceding one, include additional columns of explanatory variables and omit derived figures (averages). Each observation (an absorbance value) is in a separate row along with the values of each explanatory variable or factor. Since there are one response and four explanatory variables (7 dilutions, 2 adjuvants, 2 vaccines, and 3 tests), there will be five columns and 252 rows (7×2×2×3=84 combinations of levels and 3 replicate observations gives 84×3=252 rows). The first few rows are:

```
dilution    od     adjuvant  vaccine  test
       1  0.83470     yes       new     1
       2  0.84270     yes       new     1
       4  0.71870     yes       new     1
       8  0.59070     yes       new     1
      16  0.25170     yes       new     1
      32  0.08267     yes       new     1
      64 -0.00233     yes       new     1
       1  0.79370     yes       new     1
       2  0.79670     yes       new     1
       4  0.69570     yes       new     1
       8  0.67870     yes       new     1
       ⋮     ⋮         ⋮         ⋮      ⋮
```

Example 2 – ELISA

Incorrect

Dilution	4.0 mg/ml			Average	RP	2.3 mg/ml			Average	RP	1.1 mg/ml			Average	RP
16	0.233	0.243	0.281	0.252	7.69	0.296	0.279	0.265	0.280	3.85	0.361	0.363	0.373	0.366	1.82
32	0.389	0.329	0.314	0.344		0.429	0.426	0.421	0.425		0.612	0.594	0.590	0.599	
64	0.518	0.424	0.426	0.456		0.579	0.621	0.548	0.583		0.883	0.790	0.890	0.854	
128	0.614	0.553	0.572	0.580		0.840	0.808	0.835	0.828		1.238	1.248	1.261	1.249	
256	0.911	0.798	0.811	0.840		1.200	1.192	1.192	1.195		1.740	1.686	1.652	1.693	
512	1.215	1.148	1.151	1.171		1.747	1.708	1.616	1.690		2.038	2.079	2.168	2.095	
1024	1.763	1.637	1.535	1.645		2.037	2.040	2.047	2.041		2.450	2.354	2.339	2.381	

As in the preceding example, this display cannot serve as a raw data table because there are more than one set of observations per row (4.0 mg data, 2.3 mg data, 1.1 mg data), items that are not raw data are included (average, RP), explanatory variables appear neither in the table nor as column headings (4.0 mg/ml). In addition, cells are left blank (RP column).

Correct – Multivariate Format

To properly array these data in a multivariate format, transpose each of the columns to a separate row, include appropriate column headings, and add additional columns of the explanatory variables. The first seven fields in each row are the response vector, and the remaining fields are the explanatory variables or factors for the response vector. Explanatory variables for the individual elements of the response vector are the first seven column headings (the dilutions). The complete table is:

dilution								
16	***32***	***64***	***128***	***256***	***512***	***1024***	**concentration (mg /ml)**	**preparation**
0.23	0.38	0.51	0.61	0.91	1.21	1.76	4.0	test
0.24	0.32	0.42	0.55	0.79	1.14	1.63	4.0	test
0.28	0.31	0.42	0.57	0.81	1.15	1.53	4.0	test
0.29	0.42	0.57	0.84	1.20	1.74	2.03	2.3	test
0.27	0.42	0.62	0.80	1.19	1.70	2.04	2.3	test
0.26	0.42	0.54	0.83	1.19	1.61	2.04	2.3	test
0.36	0.61	0.88	1.23	1.74	2.03	2.45	1.1	reference
0.36	0.59	0.79	1.24	1.68	2.07	2.35	1.1	reference
0.37	0.59	0.89	1.26	1.65	2.16	2.33	1.1	reference

Example 3 – Stability Study

Incorrect

	Initial Titer			4 Month Titer			8 Month Titer			12 Month Titer			18 Month Titer		
serial	IBR	BVD	PI3	IBR	BVD	PI3	IBR	BVD	PI3	IBR	BVD	PI3	IBR	BVD	PI3
K001	3.5	6.3	4.2	3.5	6.0	3.9	3.2	5.8	4.0	3.0	–	3.8	2.9	4.8	3.1
K002	3.7	6.1	3.9	3.3	5.7	3.9	3.1	5.8	3.9	3.2	4.0	3.7	3.0	3.9	3.0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮

Correct – Multivariate Format for multiple fractions

Serial	***Month***	***IBR***	***BVD***	***PI3***
K001	0	3.5	6.3	4.2
K001	4	3.5	6.0	3.9
K001	8	3.2	5.8	4.0
K001	12	3.0	NA	3.8
K001	18	2.9	4.8	3.1
K002	0	3.7	6.1	3.9
K002	4	3.3	5.7	3.9
⋮	⋮	⋮	⋮	⋮

Example 4 – Categorical Data with Two Factors

Since there are only two factors, a two dimensional contingency table is acceptable for the submission of the data.

	normal	mild	moderate	severe	dead
Old Vaccine	14	8	3	0	1
New Vaccine	12	6	6	2	1
Placebo	2	6	4	8	11

A table in Univariate Format may be helpful if there are many levels in each factor.

Number	*Disease*	*Vaccine*
14	normal	old
8	mild	old
3	moderate	old
0	severe	old
1	dead	old
12	normal	new
6	mild	new
6	moderate	new
2	severe	new
1	dead	new
2	normal	placebo
6	mild	placebo
4	moderate	placebo
8	severe	placebo
11	dead	placebo

Example 5 – Categorical Data with More than Two Factors

Incorrect

Dairy – Nebraska

	dead	
	yes	no
Vaccine		
Placebo		

Dairy – Colorado

	dead	
	yes	no
Vaccine		
Placebo		

Beef – Nebraska

	dead	
	yes	no
Vaccine		
Placebo		

Beef – Colorado

	dead	
	yes	no
Vaccine		
Placebo		

calves (0-1 yr)

	dead	
	yes	no
Vaccine		
Placebo		

yearlings (1-2 yrs)

	dead	
	yes	no
Vaccine		
Placebo		

adult (≥ 2 yr)

	dead	
	yes	no
Vaccine		
Placebo		

In this hypothetical challenge trial, there are more than two factors. A two-dimensional contingency table representation is therefore not possible. In such cases, array the data in a univariate table.

Note also that the array of contingency tables above does not properly partition the data. Instead, the data appear twice. In the top four contingency tables, the data are broken down by site and breed, and in the lower three tables, the data are divided by age. A complete partition of the data (exclusive and exhaustive) in a univariate table will require one row for each possible combination of the levels of the five factors – treatment (vaccine, placebo), death (yes, no), breed type (dairy, beef), site (Nebraska, Colorado), and age (calf, yearling, adult). The number of rows is thus 2×2×2×2×3=48.

Correct – Univariate Table

Number	***Treatment***	***Dead***	***Breed***	***Site***	***Age***
24	vaccine	yes	dairy	NE	calf
13	placebo	yes	dairy	NE	calf
42	vaccine	no	dairy	NE	calf
⋮	⋮	⋮	⋮	⋮	⋮
19	placebo	no	beef	CO	adult

Example 6 – Field Safety Trial

Incorrect

Table 1

	number of dogs				**number of doses**				**number of events**
	< 9 weeks	**≥9 weeks**	**male**	**female**	**serial 101**	**serial 114**	**IM**	**SC**	
Site A									
Site B									
Site C									
Site D									

Table 2

Reported Event	**within 24 hours**			**after 24 hours**			**Overall by doses**	
	< 9 weeks	**≥9 weeks**	**total**	**< 9 weeks**	**≥9 weeks**	**total**	**Total**	**Rate**
inj. site pain								
swelling								
anaphylaxis								
lethargy								
vomiting								
diarrhea								
facial edema								

Do not include items which are not data (eg. total, rates). Onset (within or after 24 hours) is not an explanatory factor, but refers to the outcome. Information is not given for every combination of factors (eg. how many doses were administered IM to young puppies). There should be one row for each possible combination of the levels of the explanatory factors – site (A–D), age (less than 9 weeks, 9 weeks and older), route (intramuscular, subcutaneous), sex (male, female). The number of rows is thus 4×2×2×2=32 (and if serial were included as a factor there would be 64 rows).

The data are shown below in modified multivariate form. Columns 4-7 completely partition the explanatory factors associated with vaccine administration. The outcomes are in multivariate form in columns 9-16. Note that this is not a strict multivariate partition of the response, since a single vaccination may result in more than one of the outcome categories. Nevertheless, this modified summary format has been found useful in field safety studies. Note also that columns 2-3 are neither explanatory factors nor outcomes, but are the offset or ‘denominator’ data used in rate calculation.

The table is illustrated here broken into two linked tables, in order to fit it on the page with an easy to read font size. The key relates the unique combination of factors in the first table to the modified multivariate array of outcomes in the second. You may concattenate the tables into a single table by placing the first alongside the second.

Example 6 (continued)

Correct – Modified Multivariate Table

You may reassemble the complete table by concatenating the two linked tables. Note that the onset only subdivided one of the outcomes.

key	**doses**	**dogs**	**site**	**route**	**age**	**sex**
1	54	32	A	intramuscular	young	male
2	79	41	A	subcutaneous	young	male
3	174	96	A	intramuscular	young	female
⋮	⋮	⋮	⋮	⋮	⋮	⋮
32	213	213	D	subcutaneous	old	female

key	**pain**	**swelling**	**anaphylaxis**	**lethargy**	**vomiting**	**diarrhea < 24 hr**	**diarrhea ≥ 24 hr**	**facial edema**
1	4	1	0	0	0	1	0	2
2	2	0	1	0	0	0	0	0
3	3	2	0	4	1	0	1	0
⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮	⋮
32	1	0	1	0	0	0	0	0